Published online 1 November 2009 Nucleic Acids Research, 2010, Vol. 38, Database issue D443–D447
doi:10.1093/nar/gkp910

FlyTF: improved annotation and enhanced functionality of the *Drosophila* transcription factor database

Ulrike Pfreundt[1,2], Daniel P. James[1], Susan Tweedie[2], Derek Wilson[3], Sarah A. Teichmann[3] and Boris Adryan[1,2,*]

[1]Cambridge Systems Biology Centre, University of Cambridge, Tennis Court Road, Cambridge CB2 1QR, [2]Department of Genetics, University of Cambridge CB2 3EH and [3]Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, UK

Received August 14, 2009; Revised September 21, 2009; Accepted October 7, 2009

ABSTRACT

FlyTF (http://www.flytf.org) is a database of computationally predicted and/or experimentally verified site-specific transcription factors (TFs) in the fruit fly *Drosophila melanogaster*. The manual classification of TFs in the initial version of FlyTF that concentrated primarily on the DNA-binding characteristics of the proteins has now been extended to a more fine-grained annotation of both DNA binding and regulatory properties in the new release. Furthermore, experimental evidence from the literature was classified into a defined vocabulary, and in collaboration with FlyBase, translated into Gene Ontology (GO) annotation. While our GO annotations will also be available through FlyBase as they will be incorporated into the genes' official GO annotation in the future, the entire evidence used for classification including computational predictions and quotes from the literature can be accessed through FlyTF. The FlyTF website now builds upon the InterMine framework, which provides experimental and computational biologists with powerful search and filter functionality, list management tools and access to genomic information associated with the TFs.

INTRODUCTION

Site-specific transcription factors (TFs) are proteins that bind to specific DNA sequences or DNA conformations, and that confer regulatory information to the basal transcription machinery. While they play a key role in gene regulation in general, TFs are of special interest to developmental biologists as their presence at *cis*-regulatory elements in the genome determines important developmental decisions in processes such as axis formation and morphogenesis (1). It may therefore seem surprising that almost a decade after the availability of the *Drosophila melanogaster* genome (2) there is still no definitive answer as for the number of site-specific TFs, let alone a comprehensive list of TFs from an authoritative community resource like FlyBase (3).

FlyTF (http://www.flytf.org) has stepped in to fill this gap by integrating both computationally predicted as well as experimentally verified TFs. The first version of FlyTF (4) provided information about the curation of 1052 candidate TFs [selected for the presence of a canonical DNA-binding domain using the pipeline of the DBD transcription factor database (5) or a set of suitable Gene Ontology terms (6)], and yielded a repertoire of 753 site-specific fly TFs, about two-thirds of which were called with a high degree of confidence. The website has had ~4000 visitors since publication, with the majority of users bulk-downloading our annotations.

IMPROVED ANNOTATIONS

The initial release of FlyTF was based on *D. melanogaster* release 3.1 gene annotations, and manual curation was based on GO annotations and literature published by December 2005. The candidate proteins were primarily assessed for their capability to bind to DNA (yes/maybe/no) and confer a regulatory function. While a strict set of rules was applied for the DNA-binding property, all regulatory proteins ranging from canonical site-specific TFs to insulators and those involved in chromatin-mediated maintenance of transcription were

*To whom correspondence should be addressed. Tel: +44 1223 760209; Fax: +44 1223 760241; Email: b.adryan@gen.cam.ac.uk

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Table 1. Experimental procedures accepted to confirm DNA-binding property of candidate proteins in FlyTF, and GO terms assigned on their basis (as IDA)

Experimental procedure (literature)	Search term (FlyTF)	GO term
• Electro mobility shift assay (EMSA) • Band shift assay • Gel retardation • Low ionic strength PAGE • Sucrose gradient sedimentation	Retardation assay	GO:0003677 or GO:0043565
SELEX	SELEX	GO:0043565
Affinity chromatography	Affinity chromatography	GO:0003677 or GO:0043565
Yeast 1 hybrid screen	Y1H	GO:0043565
Yeast double interaction screen	Yeast double interaction screen	GO:0043565
• MNase digestion • DNase I footprint • Hydroxyl radical footprint	Footprinting assay	GO:0043565
Chromatin immunoprecipitation assay	ChIP	GO:0003682
Staining of polytene chromosomes	Staining of polytene chromosomes	not assigned

For a key to GO terminology, please refer to Supplementary Table S1.

treated alike. This was identified as a major limitation in computational studies that focussed on classical TFs. Furthermore, gene annotations in *D. melanogaster* are currently in their 5.19 release, meaning that many novel or modified gene models were not present in the initial dataset.

We have addressed these shortcomings in the current release of FlyTF. First, we generated a novel candidate gene list by incorporating the initial FlyTF gene set, DBD searches on the FlyBase release 5.8 gene annotations (all translations), and GO searches with a set of TF-related GO terms. This yielded a non-redundant set of 1162 candidate TFs. Two human curators (one general curator at FlyTF, one GO curator at FlyBase) assessed this list, taking all experimental evidence published by December 2008 into account.

Each candidate TF was characterised both for its DNA-binding as well as regulatory characteristics. A verdict for DNA-binding can now be

- 'yes' (clear evidence for sequence-specific binding),
- 'yes' (homolog) (property experimentally shown for a homolog),
- 'yes' (DNA binding, no sequence-specificity determined),
- 'yes (heterodimer) (if the factor alone is not capable of binding DNA),
- 'maybe' (none or no convincing evidence found) and
- 'no' (experimental evidence against DNA-binding).

As in the previous version, where available, quotations from the literature were extracted along with an associated PubMed ID. To allow users a more fine-grained selection of evidence, experiments regarding the DNA-binding characteristics of the proteins were categorised into eight different groups of varying quality, each of which can now be queried or filtered for at FlyTF (Table 1). While a DNA-binding protein in the original version automatically became a bona fide TF if the DBD pipeline identified a domain frequently found in TFs, we now provide a more detailed categorisation of the regulatory property of the candidate protein. A verdict for this can be

- 'yes' (a true site-specific TF),
- 'yes' (heterodimer) (as before, but only as a heterodimer),
- 'maybe' (if a canonical DNA-binding domain was found, but no experimental evidence) or
- 'no' (not a site-specific TF).

The 'maybe' and 'no' categories are frequently associated with free text, describing further characteristics where the information was easily accessible. Useful information in this context could be, for example, 'chromatin-remodelling', 'TBP-associated factor (TAF)', 'inhibitor' or 'insulator'. This verdict is supported by quotations from the literature as well as a discrete categorisation of the experimental evidence, which can be used for user-defined queries (Table 2).

Ultimately, in collaboration with FlyBase, any supporting experimental data was translated into GO annotation, combining the expertise of the FlyTF and FlyBase curators (the rules for the translation of experimental evidence into GO terms can be found in Tables 1 and 2). At the same time, each candidate TF received a final score based on its DNA-binding domain, and the experimental evidence found for both DNA-binding and transcriptional regulatory function (Table 3).

ENHANCED FUNCTIONALITY AND ACCESSIBILITY

The initial FlyTF website was a collection of static HTML pages and a few dynamically generated lists. A search tool to find individual genes or all TFs with a certain DNA-binding domain was the only means of user interaction. However, most visitors chose to download our annotations in bulk. We suspect this is because traditional *Drosophila* geneticists often prefer to retrieve information about 'their favourite gene' directly from FlyBase, the authoritative community resource. Also, researchers

Table 2. Experimental procedures accepted to confirm transcriptional regulatory property of candidate proteins, and evidence codes in support of GO terms dealing with regulatory function

FlyTF term	Explanation	GO evidence code for GO:0030528 or children thereof
Reporter assay *in vivo*	Any experiment that used the putative target sequence of the TF joined to a reporter gene, and showed specific activation (or repression) of the reporter through this sequence by the TF *in vivo* (embryo, larva, adult eye, etc.)	IDA
Reporter assay in cell culture	As above, but showing activity in a cell culture assay	IDA
Expression analysis	Analysis of the expression of a (or more) putative target gene(s) of the TF in mutant backgrounds (loss- or gain-of-function)	IMP
Genetic interaction analysis	Factor deemed a TF as a result of a modifier screen (not a strong evidence)	IGI
In vitro transcription assay	Transcription assay in a cell-free medium	IDA
Mutant phenotype analysis	Analysis of the phenotype after loss- or gain-of-function of the TF gene	Not assigned
Microarray	Gene expression changes as determined by microarray	Not assigned
Fusion protein with DNA binding domain followed by reporter assay	Regulatory domain of the putative TF fused to a DNA-binding domain (e.g. LexA), followed by reporter assay *in vitro*	IDA

Table 3. FlyTF score based on computational predictions (DBD) and novel GO annotation (based on experimental data)

FlyTF score	Minimal criteria (GO term and/or evidence)	Number	
1	Sequence-specific DNA binding (IDA or ISS) AND any evidence for regulation of transcription (IDA, IMP, or IGI)	133	This is analogous to previous annotations "Yes" and "maybe".
2	DNA binding (IDA or ISS) AND any evidence for regulation of transcription (IDA, IMP, or IGI)	26	
3	IDA for regulation of transcription AND assignment of a preferred DBD: homeodomain, Pax, POU, HLH, Forkhead, T-Box, Ets, bZIP, GATA, Cut, Prox1, Stat, GCM, C4 zinc finger, p53, HTH, SRF	13	
4	IDA for regulation of transcription AND assignment of any other putative DBD	10	
5	IDA for any kind of DNA binding, no experimental evidence for transcriptional regulation	110	
6	Any kind of DBD assignment (including predictions from InterPro), no evidence for transcriptional regulation.	460	
7	Little evidence for TF activity (but unlikely to be a site-specific TF)	191	
8	No evidence for any TF activity (likely to be something else)	219	

Candidate proteins in all categories can have an additional 'chromatin' call if the FlyTF curator felt the factor was more likely involved in general chromatin-related processes rather than gene-specific transcriptional regulation.

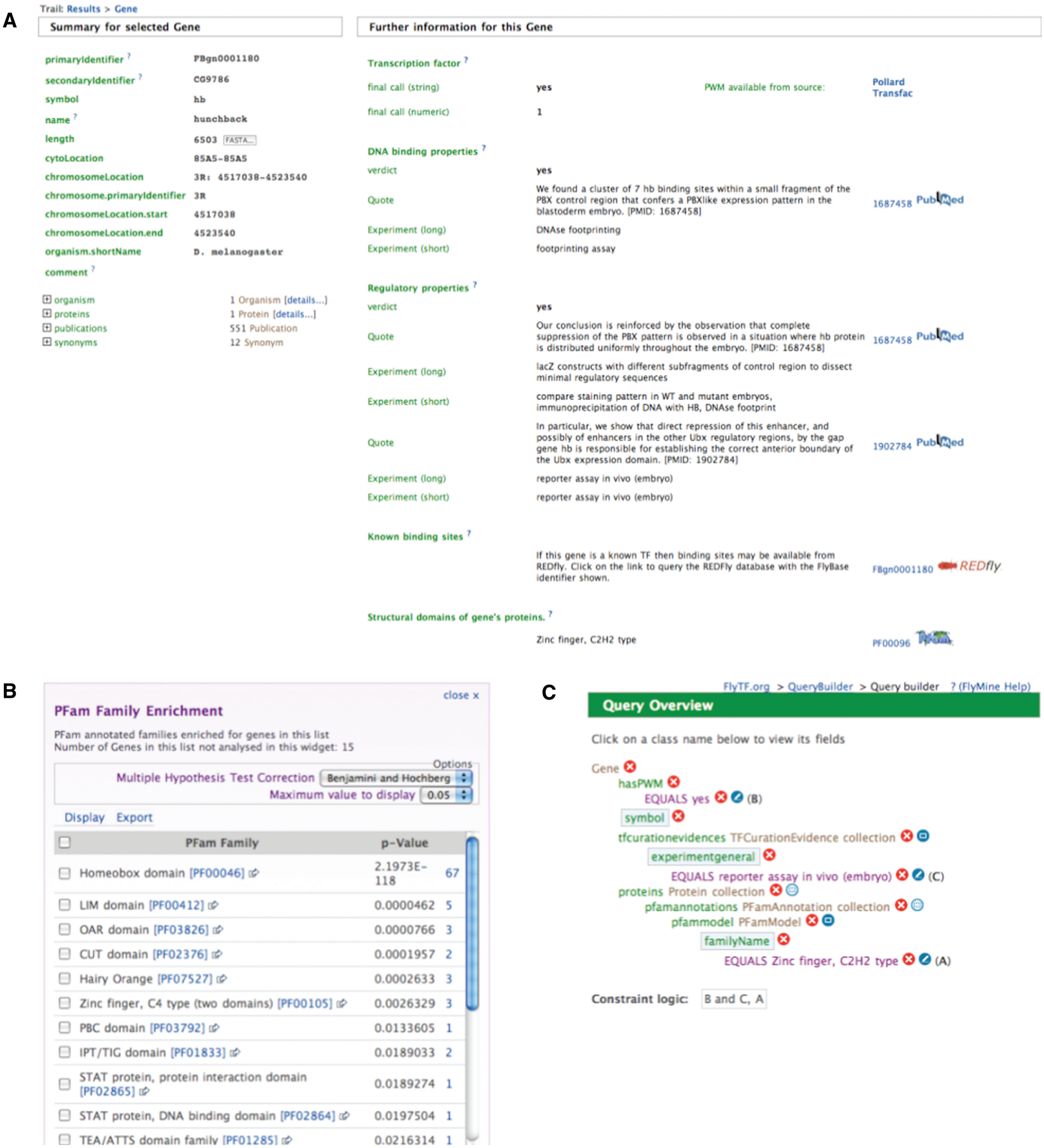
utilising genomics or computational methods are likely to query large batches of identifiers, and their analysis is often based on further list operations, neither of which were catered for by FlyTF.

We assessed a variety of options to allow non-specialist users easy access to our annotations and at the same time provide computational biologists with some basic analysis tools. The FlyTF database is now based on the InterMine framework (http://www.intermine.org), the backbone of biological data warehouses such as FlyMine (7) or modMine (8). This now enables different usage scenarios, which we will illustrate below.

The simplest scenario is the search for a single gene of interest. The query form accepts any identifier for a given TF (gene name, symbol, unique ID or even rarely used synonyms) and displays general gene information as well as our transcription factor annotations (Figure 1A).

A novel feature is of special interest for users with a genome-wide perspective: it is possible to upload extensive gene lists, from which the genes encoding TFs will be recognised and marked, and can be saved for further analysis on the website. This enables, for example, the one-step identification and characterisation of TFs contained in candidate gene lists from genomics experiments. Analysis tools available at the FlyTF website comprise 'widgets' to report GO term enrichment or over-representation of certain structural domains (Figure 1B). It is noteworthy that some of these statistics are calculated in a transcription factor background, which may be helpful in the determination of differences between sets of TFs (rather then comparing TFs against the entire genome). Users can also choose to register at the FlyTF website, and store and compare their TF lists at a later stage.

A third usage scenario addresses the needs of the computational biologists. Lists of TFs fulfilling specific criteria can easily be created using the FlyTF QueryBuilder (Figure 1C), and customisable output formats allow the swift integration of FlyTF in many bioinformatics workflows. For example, it is possible to search for all TFs that (i) contain zinc finger domains, (ii) for which a position weight matrix is known and (iii) whose transcriptional regulatory function was shown in a reporter assay in the fly. In this case, only one gene (*hunchback*) fulfils these criteria. The gene's genomic coordinates can be exported in GFF3



Trail: Results > Gene

Summary for selected Gene

primaryIdentifier ?	FBgn0001180
secondaryIdentifier ?	CG9786
symbol	hb
name ?	hunchback
length	6503 FASTA...
cytoLocation	85A5-85A5
chromosomeLocation	3R: 4517038-4523540
chromosome.primaryIdentifier	3R
chromosomeLocation.start	4517038
chromosomeLocation.end	4523540
organism.shortName	D. melanogaster
comment ?	
⊞ organism	1 Organism [details...]
⊞ proteins	1 Protein [details...]
⊞ publications	551 Publication
⊞ synonyms	12 Synonym

Further information for this Gene

Transcription factor ?

final call (string)	yes	PWM available from source:	Pollard Transfac
final call (numeric)	1		

DNA binding properties ?

verdict	yes	
Quote	We found a cluster of 7 hb binding sites within a small fragment of the PBX control region that confers a PBXlike expression pattern in the blastoderm embryo. [PMID: 1687458]	1687458 PubMed
Experiment (long)	DNAse footprinting	
Experiment (short)	footprinting assay	

Regulatory properties ?

verdict	yes	
Quote	Our conclusion is reinforced by the observation that complete suppression of the PBX pattern is observed in a situation where hb protein is distributed uniformly throughout the embryo. [PMID: 1687458]	1687458 PubMed
Experiment (long)	lacZ constructs with different subfragments of control region to dissect minimal regulatory sequences	
Experiment (short)	compare staining pattern in WT and mutant embryos, immunoprecipitation of DNA with HB, DNAse footprint	
Quote	In particular, we show that direct repression of this enhancer, and possibly of enhancers in the other Ubx regulatory regions, by the gap gene hb is responsible for establishing the correct anterior boundary of the Ubx expression domain. [PMID: 1902784]	1902784 PubMed
Experiment (long)	reporter assay in vivo (embryo)	
Experiment (short)	reporter assay in vivo (embryo)	

Known binding sites ?

If this gene is a known TF then binding sites may be available from REDfly. Click on the link to query the REDFly database with the FlyBase identifier shown. FBgn0001180 REDfly

Structural domains of gene's proteins. ?

Zinc finger, C2H2 type PF00096 Pfam

close x

PFam Family Enrichment

PFam annotated families enriched for genes in this list
Number of Genes in this list not analysed in this widget: 15

Options
Multiple Hypothesis Test Correction: Benjamini and Hochberg
Maximum value to display: 0.05

Display Export

☐	PFam Family	p-Value	
☐	Homeobox domain [PF00046]	2.1973E-118	67
☐	LIM domain [PF00412]	0.0000462	5
☐	OAR domain [PF03826]	0.0000766	3
☐	CUT domain [PF02376]	0.0001957	2
☐	Hairy Orange [PF07527]	0.0002633	3
☐	Zinc finger, C4 type (two domains) [PF00105]	0.0026329	3
☐	PBC domain [PF03792]	0.0133605	1
☐	IPT/TIG domain [PF01833]	0.0189033	2
☐	STAT protein, protein interaction domain [PF02865]	0.0189274	1
☐	STAT protein, DNA binding domain [PF02864]	0.0197504	1
☐	TEA/ATTS domain family [PF01285]	0.0216314	1

FlyTF.org > QueryBuilder > Query builder ? (FlyMine Help)

Query Overview

Click on a class name below to view its fields

```
Gene
    hasPWM
        EQUALS yes (B)
    symbol
    tfcurationevidences TFCurationEvidence collection
        experimentgeneral
            EQUALS reporter assay in vivo (embryo) (C)
    proteins Protein collection
        pfamannotations PFamAnnotation collection
            pfammodel PFamModel
                familyName
                    EQUALS Zinc finger, C2H2 type (A)
```

Constraint logic: B and C, A

Figure 1. Screenshots from the FlyTF web site. (**A**) Transcription factor summary information for gene *hunchback*. The left panel provides basic gene information and serves as a starting point for the retrieval of DNA or protein sequences. The right panel focuses on transcription factor annotation and is divided into three main sections: our general verdict, and two sections providing details on the DNA-binding and regulatory capabilities (and the associated experimental evidence thereof). Further, there is a direct link to the appropriate REDfly page, detailing transcriptional regulatory relationships for TFs where they are known. (**B**) An exemplary 'widget' for a list of transcription factors. Here, the enrichment of PFAM domain assignments for proteins of the genes in the list is shown in comparison to the rest of the genome. In the example there is a clear over-representation of the Homeobox domain. (**C**) The entire data model behind FlyTF is accessible through the QueryBuilder, allowing the definition of complicated filters for the retrieval of TF subsets. The displayed example was chosen for its relative complexity and may not be trivial for novel users to setup. However, building a query in QueryBuilder is without doubt easier than issuing the respective SQL command in a database.

format and the translations are available in FASTA format. It should be mentioned that through the customisable output generator, it is possible to export the entire FlyTF dataset as one tab-delimited file.

FUTURE DIRECTIONS

The comparative sequencing and genome annotation of closely related *Drosophila* species (9) has provided the community with the gene repertoires of a dozen flies. Experimental data for individual genes of these non-*D. melanogaster* flies is still sparse, yet researchers interested in their TFs can use FlyTF as a starting point to identify homologous proteins using the built-in orthology mapping.

The next-generation of InterMine-based databases will enable researchers to share gene lists and analysis tools across species and data mines, and we are looking forward to assist TF researchers in other model organisms with our dataset.

COLLABORATION BETWEEN TWO COMMUNITY RESOURCES

FlyTF and FlyBase both deal with the functional annotation of fly genes, and have pooled resources for this work. While FlyTF focuses on manual curation and only on TF genes, FlyBase is the community resource for all things *Drosophila*. Although the information content of each database is distinct, both use GO terms for functional annotation and a key aim of this project was to improve GO annotation consistency between these databases, based on both computational predictions and experimental evidence using the combined expertise of the TF specialists at FlyTF and the FlyBase GO curator. We believe our collaboration can be a model for many ‘niche’ databases that are maintained on a sporadic basis, which can benefit from both the experience and the resources of an established community portal.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Nick Brown and Steven Marygold at FlyBase for enabling UP to participate in this collaboration and for comments on the manuscript. They also wish to thank Goran Nenadic and Casey Bergman for the provision of computationally marked-up TF literature (10,11), and Richard Smith, Julie Sullivan and Gos Micklem for helpful comments and technical assistance in the customization of the InterMine system.

FUNDING

FlyBase grant (National Human Genome Research Institute at the US NIH P41 HG000739 to FlyBase); EPSRC MPhil studentship (to D.P.J.); Medical Research Council (to S.T. and D.W.); Royal Society University Research Fellowship (to B.A.). Funding for open access charge: The Royal Society.

Conflict of interest statement. None declared.

REFERENCES

1. Levine,M. and Davidson,E. (2005) Gene regulatory networks for development. *Proc. Natl Acad. Sci. USA*, **102**, 4936–4942.
2. Adams,M.D., Celniker,S.E., Holt,R.A., Evans,C.A., Gocayne,J.D., Amanatides,P.G., Scherer,S.E., Li,P.W., Hoskins,R.A., Galle,R.F. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.
3. Tweedie,S., Ashburner,M., Falls,K., Leyland,P., McQuilton,P., Marygold,S., Millburn,G., Osumi-Sutherland,D., Schroeder,A., Seal,R. *et al.* (2009) FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res.*, **37**, D555–D559.
4. Adryan,B. and Teichmann,S.A. (2006) FlyTF: a systematic review of site-specific transcription factors in the fruit fly *Drosophila melanogaster*. *Bioinformatics*, **22**, 1532–1533.
5. Wilson,D., Charoensawan,V., Kummerfeld,S.K. and Teichmann,S.A. (2008) DBD–taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res.*, **36**, D88–D92.
6. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
7. Lyne,R., Smith,R., Rutherford,K., Wakeling,M., Varley,A., Guillier,F., Janssens,H., Ji,W., Mclaren,P., North,P. *et al.* (2007) FlyMine: an integrated database for Drosophila and Anopheles genomics. *Genome Biol.*, **8**, R129.
8. Celniker,S., Dillon,L., Gerstein,M., Gunsalus,K., Henikoff,S., Karpen,G., Kellis,M., Lai,E., Lieb,J., MacAlpine,D. *et al.* (2009) Unlocking the secrets of the genome. *Nature*, **459**, 927–930.
9. Drosophila 12 Genomes Consortium. (2007) Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, **450**, 203–218.
10. Yang,H., Nenadic,G. and Keane,J.A. (2008) Identification of transcription factor contexts in literature using machine learning approaches. *BMC Bioinformatics*, **9(Suppl. 3)**, S11.
11. Yang,H., Keane,J., Bergman,C.M. and Nenadic,G. (2009) Assigning roles to protein mentions: the case of transcription factors. *J. Biomed. Inform.*, **42**, 887–894.